PHYSICAL NAVIGATION OF VIRTUAL TIMBRE SPACES WITH TIMBREID AND DILIB

William Brent

American University
Audio Technology Program
4400 Massachusetts Ave NW
Washington DC

ABSTRACT

This paper summarizes recent development of two open source software libraries that enable auditory display in Pure Data (Pd), and describes developing projects that were achieved using the two packages in tandem. The timbreID feature extraction and classification library enables real- and non-real-time audio analysis via high-level modules that can be programmed for a variety of purposes. DILib (the Digital Instrument Library) provides software tools for accessing and managing gestural control streams as captured by inexpensive, widely available sensor hardware. Realized at the intersection of these software packages, three applications are discussed from technological and performative viewpoints: a system for navigating visual timbre spaces with gestures drawn from full body tracking, a similar system based on open-air infrared fingertip tracking, and the *Gesturally Extended Piano*—an augmented instrument controller that uses piano performance gestures to create visually explicit action-sound relationships.

1. INTRODUCTION

Among available music information retrieval software packages (e.g., [1][2][3][4]), those designed for use in real-time multimedia programming environments are especially valuable for visually-based audio browsing and the performance of live computer music. Such software allows artists to analyze, organize, and reshape immense collections of digitally stored sound with sophistication and relative ease. Parallel to this development—as interest in embodied computer music practices continues to grow—tools that enable the high-speed capture of body movement data have also reached a high level of refinement.

This paper begins by summarizing recent development of two software libraries for Pure Data (Pd) [5], a popular open source multimedia programming environment. The timbreID audio feature extraction and classification library enables real- and non-real-time audio analysis via high-level modules that can be programmed for use in a variety of contexts. Provided example applications include real-time speech recognition, instrument identification, target-based granular synthesis, and various types of sound visualization. The Digital Instrument Library (DILib) provides software tools for accessing and managing gesturally-oriented control streams as captured by increasingly sophisticated yet inexpensive sensor hardware. These include accelerometers, multi-touch surfaces, body tracking systems, and high frame rate digital cameras that can be used for a number of computer vision strategies.

The concerns of these two projects are distinct, but a spectrum of applications exists at their intersection that encompasses purely research-oriented sound exploration tools as well as full-fledged musical instruments. Use of physical gesture information beyond that offered by standard computer input devices enhances applications along this spectrum considerably, making it possible to achieve customized multi-modal relationships with audio based on sound, sight, and touch. The final section of this paper describes three developing projects that explore these possibilities in Pd using timbreID and DILib. Both libraries have been released under the GNU GPL as open source projects, with the intention of promoting novel modes of sound exploration and digital music performance based on freely designed action-sound relationships.

2. AUDITORY DISPLAY WITH TIMBREID

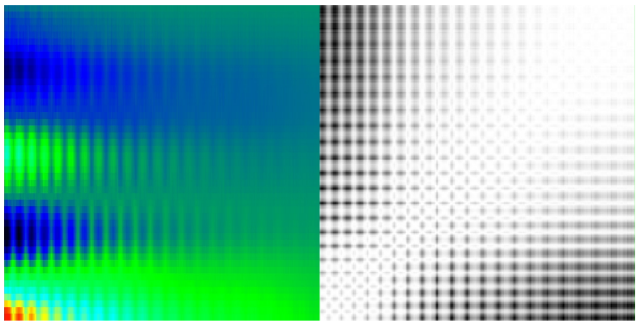

Figure 1: A Bark-frequency cepstrogram (left) and similarity matrix (right) of a tam-tam strike.

Originally described in [6], components of the timbreID library can be used for many different purposes. The current release features improvements and additions to the core analysis and data management objects as well as to the accompanying examples package. Here, we will only summarize example applications that are directly useful for auditory display, with the most significant items being spectrogram, cepstrogram, and similarity matrix plotting tools, and improved functionality of the timbre space plotter. Figure 1 shows a Bark-frequency cepstrogram and similarity matrix of a tam-tam strike that were generated using these tools. Mel- and Bark-frequency cepstrum remain popular as compact descriptors of timbre, but the choice of an optimal range of coefficients for identification tasks requires judgment based on the particular sounds and circumstances. Visualization of cepstral information in the form of a cepstrogram is useful for understanding how individual coefficients vary over the course of specific sounds, and can be a valuable aid in making these kinds of choices.

Plotting segments of audio in relation to their quantifiable features is another technique for understanding relationships between sounds, as well as for designing large and small scale sound sequences based on timbre. In this type of plot, points can be made to represent audio segments of a fixed grain size or entire sound events, and can be auditioned by moving a cursor within range. Figure 2 shows this tool as realized using timbreID, with a collection of piano samples as the objects of analysis. Grains of audio are spaced along the horizontal and vertical axes according to amplitude and spectral centroid, respectively. The axes of the plot can be chosen based on available audio features, and all feature data is displayed for the most recently browsed grain in the information panel shown on the left. Individual audio features can also be plotted against time to reveal dimensions of timbre relative to small and large scale temporal structure.

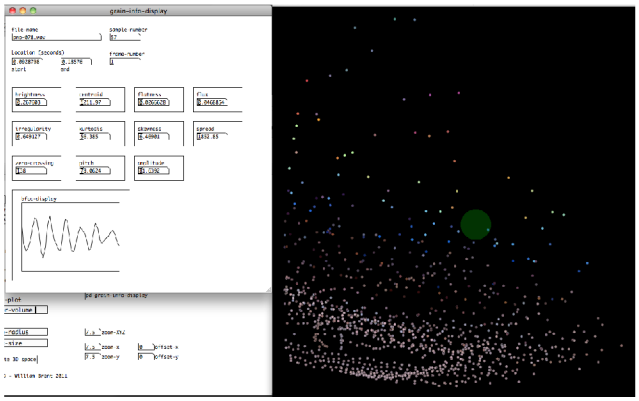

Figure 2: Timbre space plot of piano audio grains.

The main limitation of this basic plotting system is connected with dimensionality. Plots can be viewed and rotated in three dimensions, but only navigated and auditioned in two dimensions with a standard computer mouse. Considering the multidimensional nature of timbre, this is a very significant shortcoming. One alternative is to navigate the space based on the qualities of sounds captured by a microphone in real-time. By harvesting the first three Bark-frequency cepstral coefficients (BFCCs) of a live signal as it changes over time, the input sound can be used as a type of cursor moving in three dimensional space. Additional BFCCs can be used to further increase dimensionality, but attempting to make changes in any one dimension by altering the timbre of the input sound does not result in a high degree of control. Further, even in three dimensions, the process of navigating in this manner is very difficult to conceptualize visually. For better results, we need access to control streams from gesture input systems more sophisticated than the standard computer mouse.

3. GESTURE ACQUISITION WITH DILIB

The experience of using interactive sound visualization systems changes fundamentally when different types of body movement are introduced as sources of control. Research in the field of Human Computer Interaction (HCI) has yielded many robust options for capturing physical movement information with minimal encumbrance. The associated hardware and software are increasingly accessible for use within flexible environments like Pure Data, a situation that has encouraged widespread artistic application of these techniques. Moving beyond basic access, a Pure Data library is needed for parsing/routing data streams and generating additional higher level features based on raw tracking information. DILib (originally presented in [7]) aims to meet this demand.

DILib accounts for many different sources of gestural control data. Most relevant to the discussion here are those based on infrared (IR) blob tracking and full body tracking. IR blob tracking has been used as a reliable means of capturing motion information in a variety of contexts. The basic method is to shine a particular wavelength of IR light on a scene, and place highly reflective markers on key points of a moving body. Near the light source, a camera fitted with a bandpass filter tuned to the same IR wavelength observes the scene. Frames in the digital video stream are then subjected to some basic pre-processing before being fed to a blob tracking algorithm. After these steps, objects reflecting a relatively high amount of IR light back to the camera will appear in the video stream as white blobs, while less reflective objects are rendered completely black. Thus, motion within a diverse scene can be reduced to just a few key points of interest.

A significant problem associated with this technique has to do with distinguishing between the tracked blobs. To overcome this, some type of history and analysis of the blob trajectories must be maintained in software. DILib's IR blob tracking module was built using objects in the Graphics Environment for Multimedia [8], and core DILib objects for managing blob continuity and extracting higher level features from blob position data. These features include distances, angles, and centroids between pairs of points, and delta values of individual points across frames. Specific gestures (e.g., pinching and rotation with the fingertips) can be identified based on these features in order to offer different classes of control over synthesis and spatial navigation.

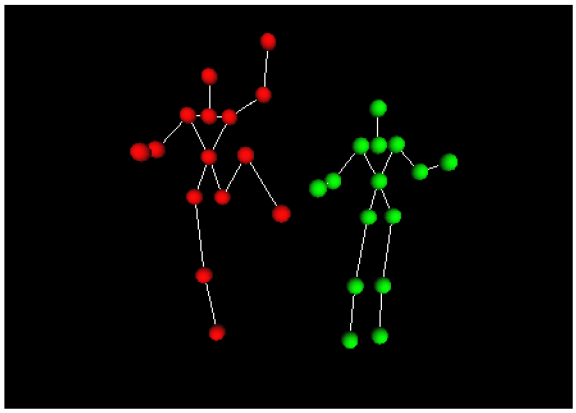

Figure 3: OSCeleton frame data rendered in Pd/GEM with DILib

More sophisticated feature analysis is required for full body tracking, where raw sensor information must be interpreted relative to a model of human movement. DILib's module for body tracking relies entirely on external software for this fundamental step. OSCeleton[1] is open source multi-platform software that interprets data from Microsoft's Kinect sensor and produces three-dimensional coordinates for the primary points of a body being tracked. Its output can be received in Pd via OSC messages, where DILib offers objects for managing data streams of multiple users,

[1] https://github.com/Sensebloom/OSCeleton

graphical rendering of the skeleton frame (shown in Figure 3), and generation of relative data (e.g., distances between extremities, angles at the elbows and knees, etc.).

An important variety of relative data is the offset of an extremity from its attaching joint, such as the three-dimensional position of the right hand in relation to the right shoulder as an origin. Using this approach, the raw coordinate of a user's hand in the entire scene can be polled to control global aspects of a system, while its offset from the shoulder maintains a high degree of independence and is suitable for control over more specific aspects. In this body tracking module and more generally, a central aim of DILib is to facilitate the design of systems that produce complex but consistent consequences in response to changes in basic sources of data. As with acoustic instruments, such systems present an interesting set of constraints, where individual parameters can be modified with near—but not complete—independence.

4. APPLICATIONS

This section reviews characteristics of three real-time sound exploration/instrument systems. In all cases, a fundamental concern is the pursuit of methods for translating continuous movement data into continuous changes in timbre. A second (and somewhat contradictory) function is transformation of this core sound via a layer of dynamically routed signal processing modules.

4.1. Embodied Timbre Space Navigation

In the context of a timbre space, the skeleton frame data described in Section 3 can be used in any number of ways. In the simplest case, a subset of the skeleton's primary points can be used as three-dimensional browsing/auditioning cursors. This has the immediate benefit of providing polyphony, making it possible to reach toward multiple timbre regions at once, and pushes a basic exploration tool closer to becoming a musical instrument. A fundamental property of digital musical instruments is their ability to dynamically reassign pre-defined action-sound relationships (i.e., mappings), and here, nothing restricts the implementation of several different strategies that can be chosen freely during use.

The current system offers three navigation environments, which can be chosen by walking through one of three virtual "doorways" at a specific depth threshold within the physical tracking area. From the extreme rear of the tracking area and facing the sensor, walking forward to cross the depth threshold at the leftmost region imposes the simple multi-cursor mapping described above. The left and right hand are designated as active cursors, while distance between the hands and their individual three-dimensional delta values (i.e., accelerations) modulate parameters of various processing modules. Traversed in the other direction, the depth threshold is used to deactivate the mapping, freeing the user to cross it again at either the center or rightmost regions.

Mappings in the remaining doorways explore possibilities that arise when timbre spaces are grafted directly on the shoulders of the user. That is, rather than spreading audio grains throughout the entire tracking area, they are compressed to cubes attached to the users shoulders and auditioned based on the relative offset of the corresponding arm. Under this approach, a specific arm gesture activates roughly the same sequence of grains regardless of where the user stands in the tracking area. This means that the user's overall position can be used to select different chains of signal processing for application to the basic granular output. Leaning into specific regions, the user can choose to apply a network of flanging, pitch shifting, and pulsing at one moment, but ring modulation, filtering, and reverberation at the next. This embodied approach to timbre space navigation and audio processing provides access to a greater number of options, varies the orientation between user and space, and generally enhances large scale physical aspects of interacting with digital audio.

4.2. Open-air Fingertip Navigation

More nuanced control can be attained by browsing timbre spaces via open-air fingertip movements. Technically, this system relies on IR blob tracking, with reflective markers placed on tips of the thumb and middle finger of each hand. Because the markers are lightweight and passive (i.e., not powered), movement is not restricted. IR motion capture systems typically involve multiple cameras in order to capture data with three degrees of freedom. Here, the system is drastically reduced in comparison because the tracking area is relatively small, and portability, cost, and ease of use are top priorities. Nevertheless, it does provide very reliable tracking, including excellent depth resolution for three-dimensional tracking. Without additional cameras, spherical markers (which appear to be the same size from any angle at a given distance) are required in order to use IR blob size as an indicator of depth.

Rather than virtual doorways, pre-defined mappings are chosen based on which of the four fingertips enters a particular side of the tracking area first. A similar strategy was used effectively for an instrument described in [9]. As before, relative data between points can be used to modulate parameters of processing applied to the audio grains as they are browsed. For instance, by pinching with the left hand and rotating the wrist, the user can make specific adjustments to variables like delay time and pitch shift interval. The shape and size of the polygon defined by the four fingertips can be used for other layers of control. Considering the system as an instrument, we can say that its sound producing actions are extremely indirect, happening in relation to virtual objects that the performer must see to understand. With practice, strong relationships are formed between visual characteristics of the virtual elements and the resulting audio output.

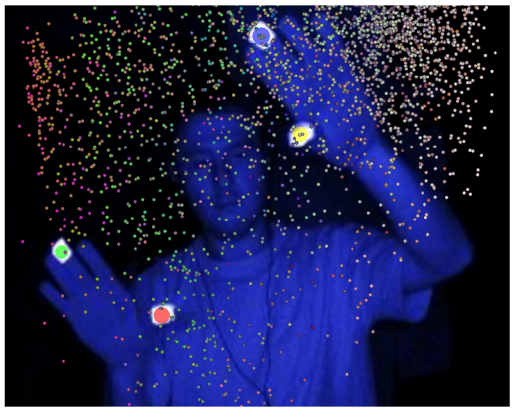

Figure 4: IR fingertip tracking for polyphonic timbre space browsing.

4.3. The Gesturally Extended Piano

The Gesturally Extended Piano (GEP) is an augmented instrument controller that exploits the pianist's arm movements for timbre space navigation and control over real-time transformation of the piano's acoustic sound. Among the most elementary pieces of movement information in the case of a pianist are the positions and angles of the forearms in relation to the keyboard. This information can be captured with IR blob tracking by following a minimum of two key points on each arm. As well as allowing different timbre spaces to be grafted onto the specified region of interest for polyphonic browsing with the four reflective markers as cursors, augmenting the piano with motion tracking enables intuitive control over sound characteristics that are usually inaccessible when playing the piano, such as continuous changes in pitch and volume.

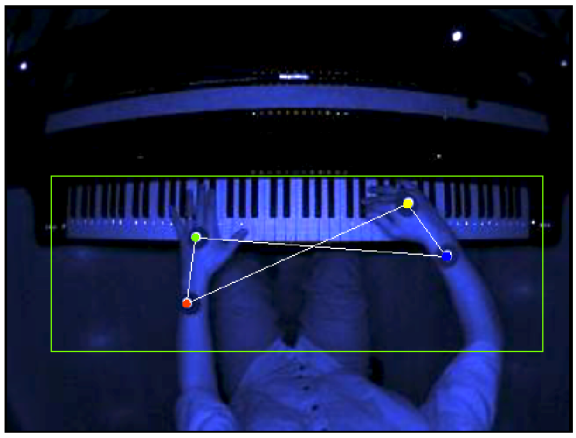

Figure 5: Overhead view of the GEP tracking system.

The GEP's camera and attached IR light array should be mounted directly over the keyboard in order to provide a clear overhead view of the entire playing surface and the pianist's arms. A convenient mounting point on grand pianos is the raised lid, but upright pianos can be fitted with the tracking system as well. The spherical reflective markers should be attached to the pianist's arms using a flexible silicone skin adhesive. Figure 5 shows the IR camera's view of the piano and user-defined region of interest, with red, green, blue, and yellow points drawn over top of the reflective markers, and connections drawn between some points. This animation provides useful feedback for the performer, and (as with the other systems described in this section) several interdependent control streams can be extracted from the scene.

Different mapping presets for the GEP controller can be selected based on entry conditions of the hands. For instance, the hands can enter from either the middle, far left, or far right of the region of interest, which provides three preset choices. The number of available choices can be doubled by observing whether the right or left hand is the first to enter each of these zones. Based on a depth threshold, the number of choices can be doubled once again, meaning that the pianist can choose to enter the region of interest either above or below the invisible threshold. This strategy avoids the need for any additional pedals or switches, keeping the amount of hardware to a minimum.

Space does not permit a detailed explanation of the mappings currently in use; however, one of the more intriguing options involves phase-vocoded scrubbing of a short audio buffer filled incrementally with a mix of desired audio fragments. This mapping relies on the distance between points on each hand, which can be lengthened or shortened by flexing the wrist forward or back. By defining a threshold, these motions can be used to trigger live audio capture into the buffer with the left hand, and clearing of the buffer with the right. The pianist can thus trigger the left hand before playing into the buffer, which is then scrubbed using the centroid of all four tracked points. Moving the hands between the low and high extremes of the keyboard, any particular moment of the sampled sound can be sustained by virtue of the phase vocoder, with further processing controlled via other aspects of arm orientation. After building up such a texture incrementally, the buffer clearing trigger of the right hand provides a means of bringing dense, sustained sound masses to a sudden and dramatic halt.

5. CONCLUSION

Both timbreID and DILib have been released under the GNU GPL as open source projects with the intention of further encouraging embodied approaches to digital exploration of sound relative to timbre. Though designed for native use in Pd, information generated by these libraries can be routed to any multimedia programming environment. Of the specific applications reviewed in Section 4, only the GEP has been used in live performance. After a period of experimentation, use, and refinement, software for these projects will be made available as open source tools for interested artists and performers.

6. REFERENCES

[1] G. Tzanetakis and P. Cook, "Marsyas: a framework for audio analysis," *Organised Sound*, vol. 4, no. 3, pp. 169–175, 1999.

[2] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, France, 2007.

[3] J. Bullock, "Libxtract: A lightweight library for audio feature extraction," in *Proceedings of the International Computer Music Conference*, 2007.

[4] N. Collins, "SCMIR: A SuperCollider music information retrieval library," in *Proceedings of the 2011 International Computer Music Conference*, 2011, pp. 499–502.

[5] M. Puckette, "Pure data: Another integrated computer music environment," in *The 2nd InterCollege Computer Music Concerts*, 1996, pp. 37–41.

[6] Author, "A timbre analysis and classification toolkit for pure data," in *Proceedings of the International Computer Music Conference*, 2010, pp. 224–229.

[7] ——, "DILib: Control data parsing for digital musical instrument design," in *Proceedings of the 4th International Pure Data Convention*, 2011, pp. 176–180.

[8] M. Danks, "Real-time image and video processing in GEM," in *Proceedings of the 1997 International Computer Music Conference*, 1997, pp. 220–223.

[9] J. Oliver, "The MANO controller: A video based hand tracking system," in *Proceedings of the 2010 International Computer Music Conference*, 2010.